Vision and Strategy for a DOE Science Grid

William E. Johnston
Distributed Systems Department
National Energy Research Scientific Computing Division
and
Information Power Grid Project,
NAS Division, NASA Ames Research Center

The Vision for a DOE Science Grid

Large-scale science and engineering is typically done through the interaction of

- **people,**
- **heterogeneous computing resources,**
- **multiple information systems, and**
- **instruments,**

all of which are geographically and organizationally dispersed.

The overall motivation for "Grids" ([2],[3]) is to *enable the routine interactions* of these resources to facilitate this type of large-scale science and engineering.

Two Sets of Goals

Our overall goal is to facilitate the establishment of a DOE Science Grid **("DSG") that ultimately incorporates production resources and involves most, if not all, of the DoE Labs and their partners.**

A "local" goal is to use the Grid framework to motivate the R&D agenda of the LBNL Computing Sciences, Distributed Systems Department ("DSD").

Applications

Several types of science and engineering scenarios **are driving the development and deployment of Grids at DOE and NASA:**

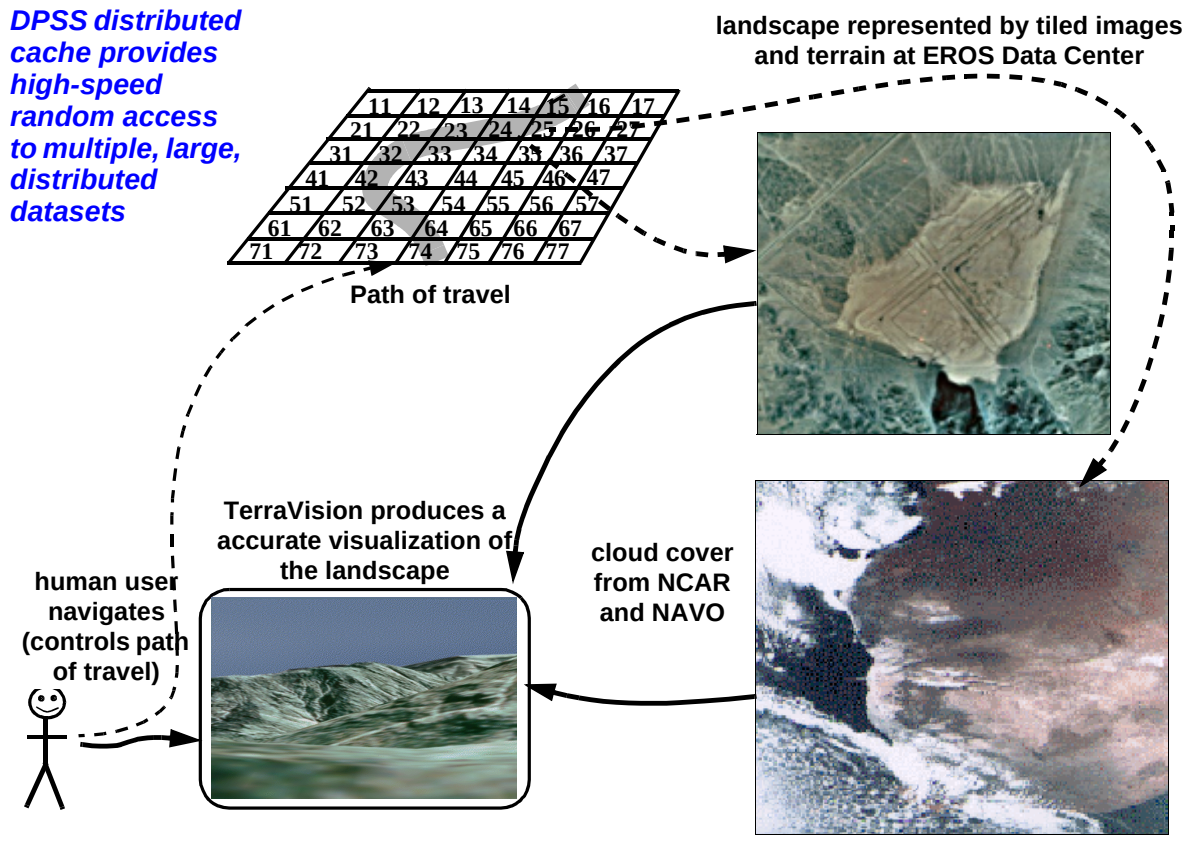
- ***Large-scale, multi-institutional engineering design*** **and multi-disciplinary science – e.g., design of next generation diesel engines, next generation space shuttle, etc.**
- ***Scientific data analysis and computational modeling with a world-wide scope of participants*** **– e.g. High Energy Physics data analysis**

- ***Real-time data analysis for on-line instruments***, **especially those that are unique national resources – e.g. LBNL's and ANL's synchrotron light sources, PNNL's gigahertz NMR machines, etc.**
- ***Coupling of laboratory instrument experiments*** **and computational models to support, e.g., experiment and computational steering**
- ***Generation and management of large, complex data archives*** **that are shared across an entire community – e.g. DOE's human genome data and NASA's EOS data**

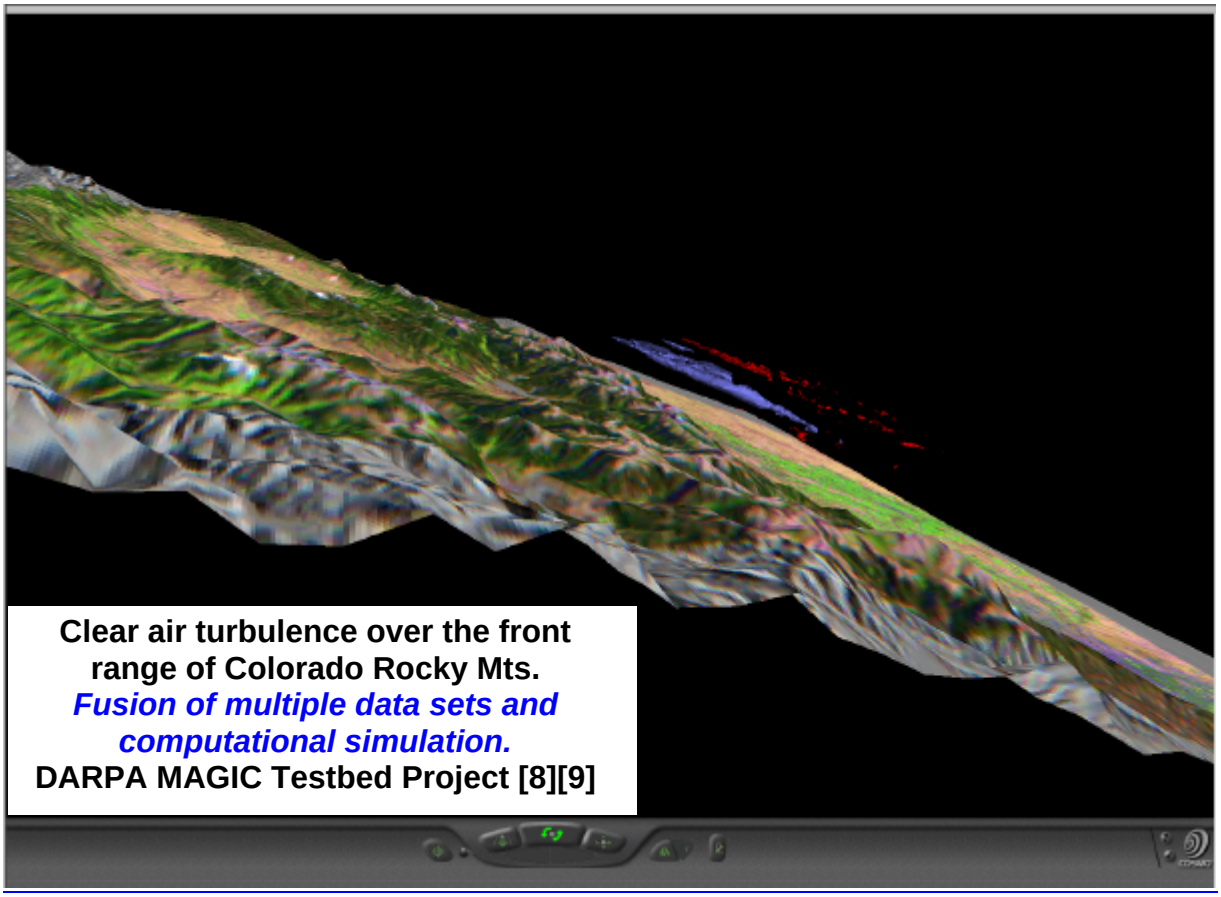
Examples of Grid-like Systems

- ***High data-rate, widely distributed data management*** **(federated access for archived satellite and aerial imagery, digital terrain data, and atmospheric data in the MAGIC Gigabit Testbed [8], [9])**
  - **on demand, real-time interactive exploration of an operational environment supporting, e.g., military operations and community emergency services**
  - **aggregation of multiple, widely distributed, multi-discipline data sets**

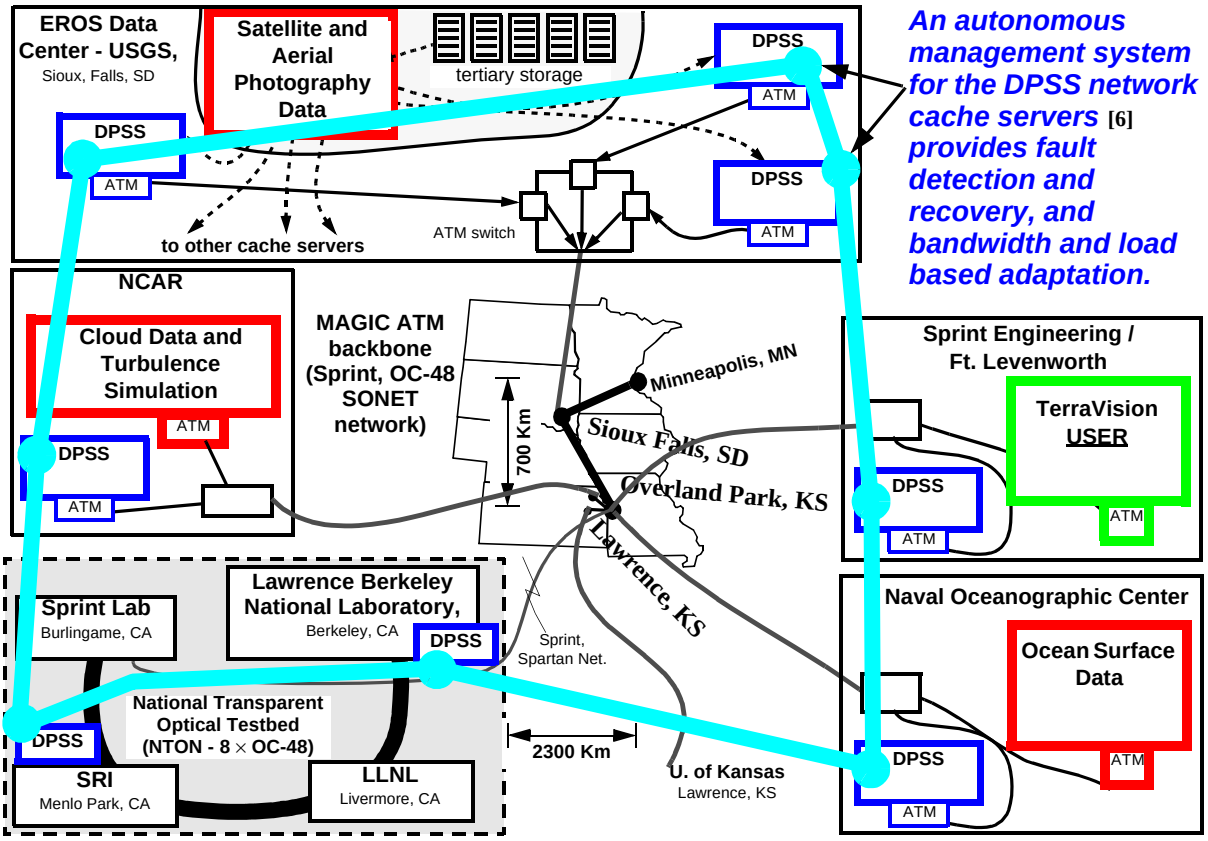
- on-line, real-time access to multiple environmental data sets that are (and always will be) maintained by domain experts at their own sites.
- DARPA MAGIC testbed consortium (see www.magic.net) developed distributed services, data and visualization from EROS Data Center, NCAR, NAVO, SRI
- *Similar characteristics to DOE NGI applications like the Combustion Corridor [5]*



TerraVision Provides Real-time Visualization of Aggregated Data



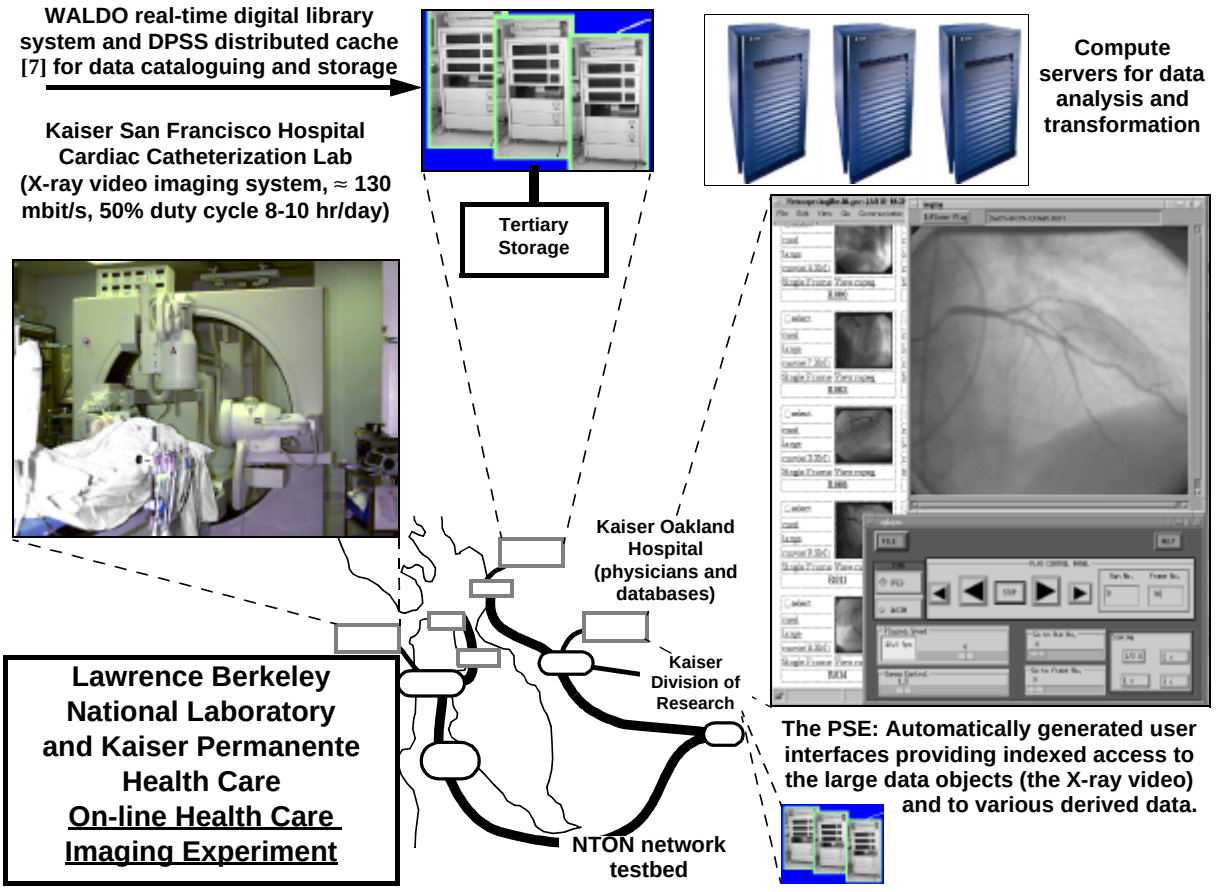
Clear air turbulence over the front range of Colorado Rocky Mts.
Fusion of multiple data sets and computational simulation.
DARPA MAGIC Testbed Project [8][9]

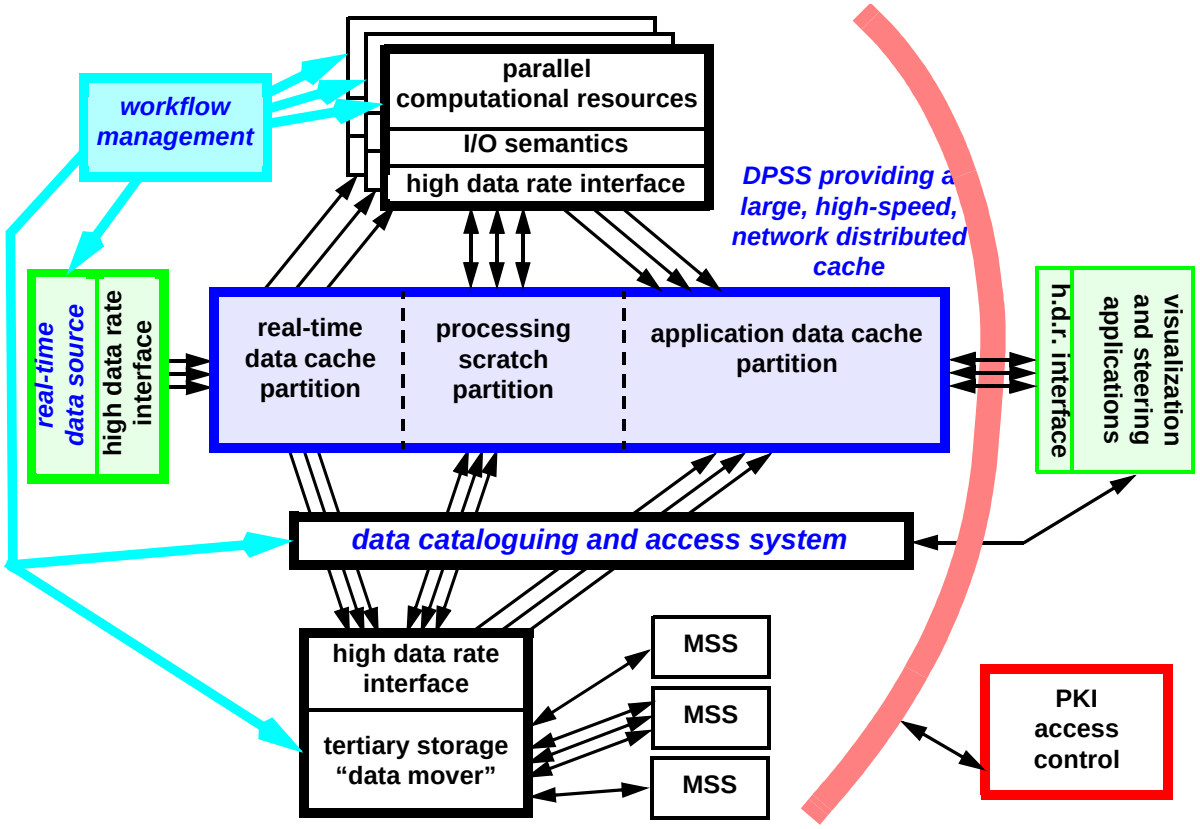


An autonomous management system for the DPSS network cache servers **[6]** ***provides fault detection and recovery, and bandwidth and load based adaptation.***

The MAGIC Testbed Distributed Application Environment

- **On-line medical imaging system (*real-time digital libraries for on-line, high data-rate instruments* [7])**
  - **on-line, real-time, high data-rate medical instrument with remote users**
  - **distributed data analysis and automatic data cataloguing and archiving**
  - **strict authorization and access control**
  - **optical WDM metropolitan area network (NTON)**
  - ***Similar characteristics to DOE NGI projects like Clipper [11] and the Physics Particle Data Grid [5]***



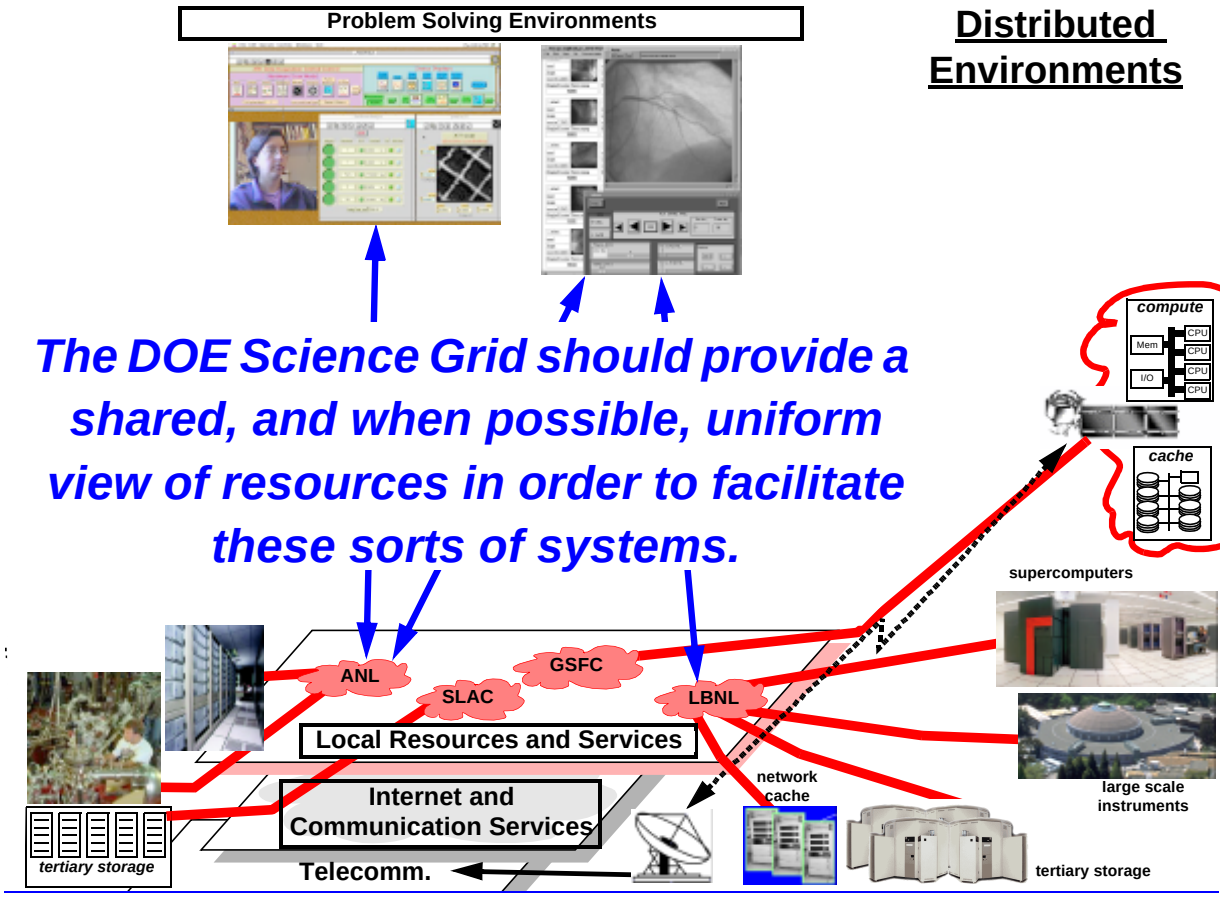


A High Volume, High Data Rate, Data Analysis Architecture

These large-scale science and engineering problems involve many types of applications and data sources that are accessed and shared across many institutions. This implies:

- **numerous interconnected servers providing computational simulation and analysis, data access, and functional access to instruments, in semi-open agency/research networks (e.g., ESNet, NREN, Internet-2, etc.)**
- **many simultaneous collaborators, e.g. at DOE Labs, NASA Centers, other Federal labs, industrial partners, and universities**
- **many stakeholders and diverse assets .**

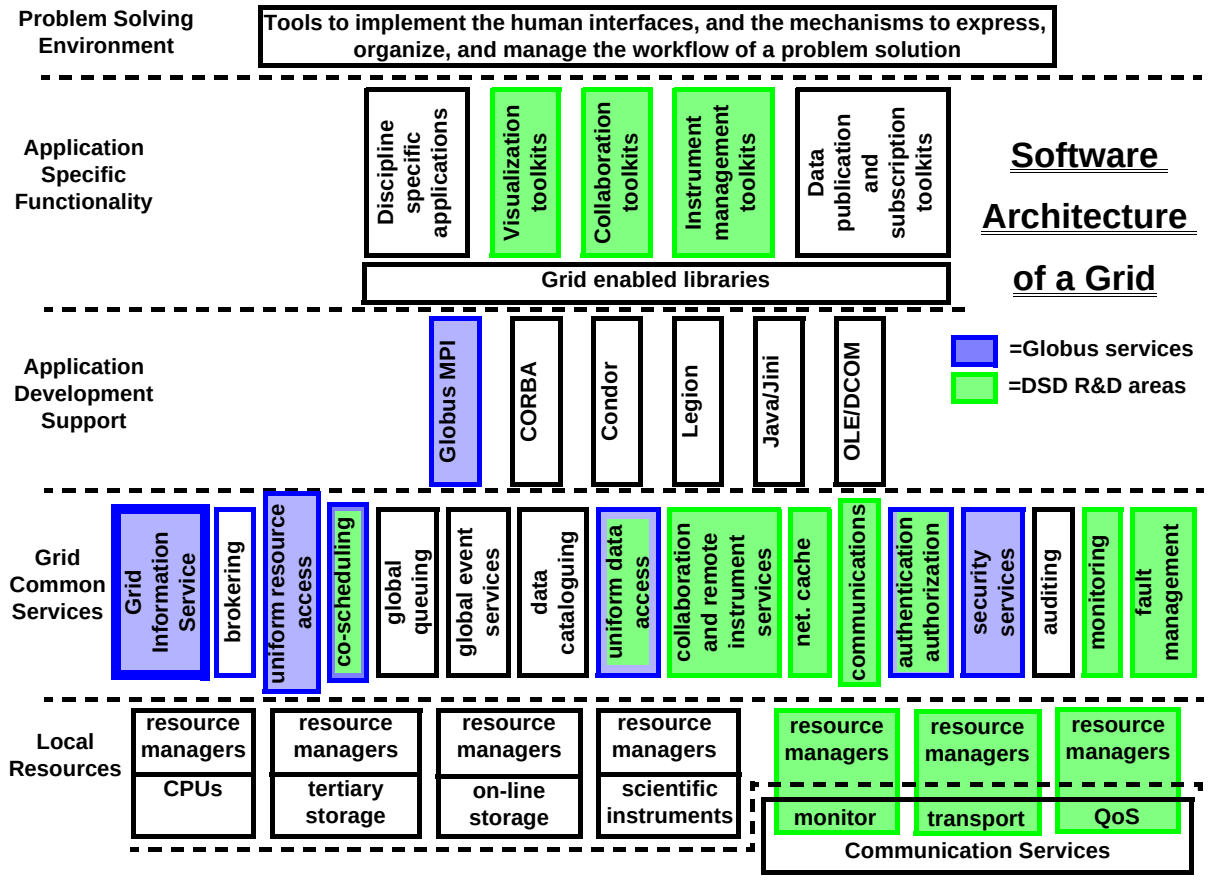
Distributed Environments



The DOE Science Grid should provide a shared, and when possible, uniform view of resources in order to facilitate these sorts of systems.

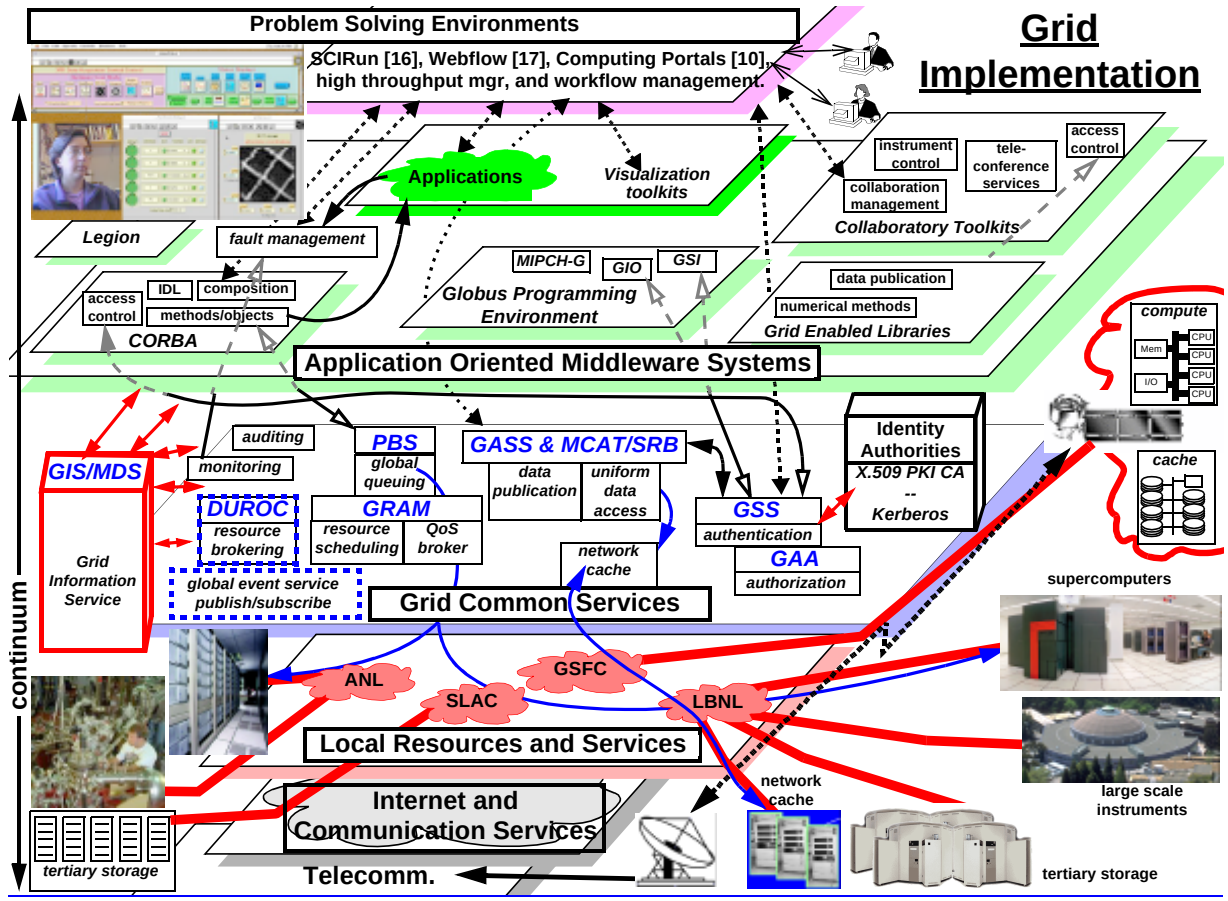
Grids

- ***Grids will provide a consistent, persistent, and supported distributed computing, data, and collaboration environment.***
- ***Grids are built through collaborative efforts*, and at the same time facilitate collaboration.**
- ***Grids should provide a rich set of architecturally consistent services* for constructing, using, and managing the types of diverse, widely distributed environments described in the preceding examples.**



Grids

- ***These Grid services will be used to organize, aggregate, and share diverse resources to provide highly capable and uniform access*** **to widely distributed computing, data, instrument, and human resources at DOE Labs around the country.**



Grids

- ***To be useful for applications, the DOE Science Grid must also provide operational support*** **and persistent infrastructure, together with various sorts of user support.**
- ***For such an open environment to be feasible, security must be a design goal from the start*****, and must address authentication, authorization, and infrastructure assurance**

***This environment will enable applications to routinely use widely distributed resources*. E.g:**

- ***applications that require locating and co-scheduling many resources***
- ***configurable problem solving environments for very large parameter space studies***
- ***on-line instruments coupled to schedulable computational and data resources***
- ***collaborative interaction with all resources*, including, e.g., multi-party analysis and visualization of massive datasets**

WEJ Goals for the DOE Science Grid

- ***Supported and persistent Grid resources*, technology, and services within the ESNet and NERSC production environments, and across the DOE Labs**
- ***Applications that test Grids and benefit from Grids***
- ***An R&D program addressing open issues in Grid* computing and data management that are revealed in the course of building and operating the DSG**
- ***Establishment of the production DSG* as the service delivery model for DOE computing, data, and instrument resources**
- ***Active participation in the Grid Forum* [12]**

DOE Science Grid Strategy

- ***Collaboration among DOE Labs*** **and active cooperation with the NASA and NSF Centers**
- ***Build and support a DOE Science Grid testbed*** **– initially across LBNL, ANL, and SNL – based on the groundwork of DOE's NGI testbed, Globus [1], NASA's IPG [3][4], etc.**
- ***Promote and study application use of DSG testbe*****d**
- ***Evolve the testbed into a prototype production environment*** **with the participation of NERSC and ESNet**

The Importance of a Persistent Grid Infrastructure

"Persistent" means that Grid services are always available on a significant and stable set of computing and storage resources.

- ***Application developers will not put the effort into learning how to use distributed environments*** **until they are convinced that there is a persistent and supported approach to dealing with distributed resources**
- ***Application communities will not plan future capabilities based on use of distributed resources*** **unless they are convinced that a persistent and supported infrastructure will exist**

- ***Operational organizations will not develop the approaches to reliably operate*** **a distributed environment until it is clear that the basic structure of these environments is defined and being built**
- ***The Grid R&D community will not learn the limitations of the current approaches*** **until they are put into place and exercised by a wide range of application communities**

  (and each new agency - NASA, DOE, NIH - that embraces Grids will bring a new set of problems and applications that will force development of new services and operational procedures to accommodate those new application areas and communities)

DOE Science Grid Testbed Initial Approach

- ***Grid services***
  - **Globus providing basic Grid common services [1]**
  - **X.509 identity Certification Authority**
  - **Grid Information Service (Globus GIS/MDS)**
  - **system admin, account mgmt., and user support**
  - **global queuing mechanism (PBS++ [13])**
  - **job tracking and monitoring**
  - **access to tertiary storage (GASS, MCAT/SRB [14])**
  - **programming and program execution support (MPICH-G, CORBA, and Condor [15])**
  - **network bandwidth reservation**

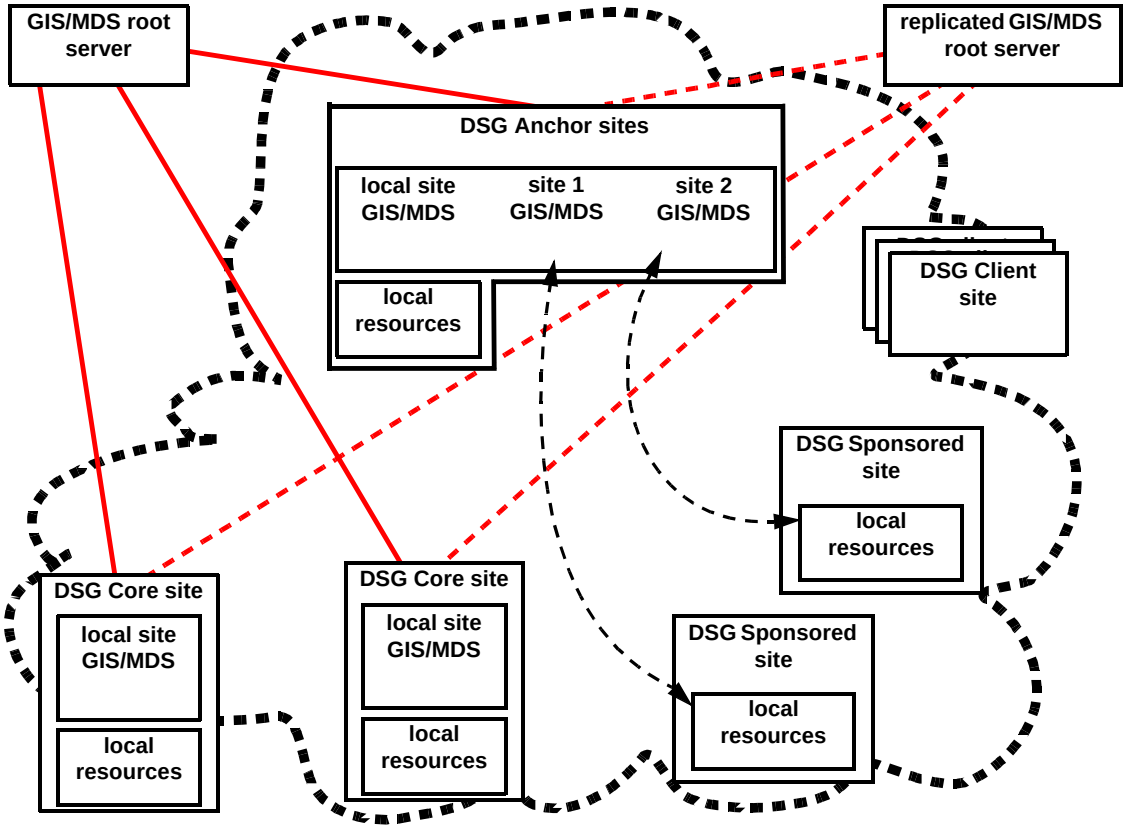
¨ ***Operational model*** **(based on experience at NASA)**

- **Anchor sites: For each major Grid (e.g. DOE Science Grid) there will be "anchor" sites that provide persistent production services. NAS plays this role for NASA/IPG. ESNet and NERSC will likely play this role for DOE/Science Grid.**

A "Grid" consists of resources federated through compatible Grid Common Services and the use of a common root name in the Grid Information Service. Whether the federated resources are actually used together depends on the existence of user allocation and/or fungibility of allocations, acceptance of a common X.509 identity policy, adequate network bandwidth, etc. Collections of resources associated by a common funding agency/program (e.g. DOE/Science/MICS) are likely to have a common user community. Sites may operate independent Grid services. Anchor sites may provide things like 7x24 operations, trouble ticket routing, etc.

NERSC and ESNet will fill somewhat different roles as anchor sites: NERSC for its services and user base, and ESNet more generally (and probably with a smaller range of services) for DOE/Science Grid.

- **Core sites: Sites like GRC and LaRC in IPG and likely LBNL, ANL, PNNL, ORNL, ACL, etc. in DOE Sci. will operate their own distributed portions of the infrastructure.**
- **Sponsored sites: Sites that run the Grid Common Services / Globus on their local systems, but defer to the anchor sites to run their GIS/MDS, CA, provide user support, etc.**
- **Client sites: Sites whose participation is limited to users obtaining an appropriate X.509 cert. and installing the client software so that they can access Grid services. (And things must be made easier for these folks!)**



Operational Model for DOE Science Grid Information Service

DOE Science Grid Testbed

- ***Incorporate persistent resources***

  Resources, together with suitable local resource management systems, must be sufficient to support and initial set of applications (e.g some DOE NGI applications)

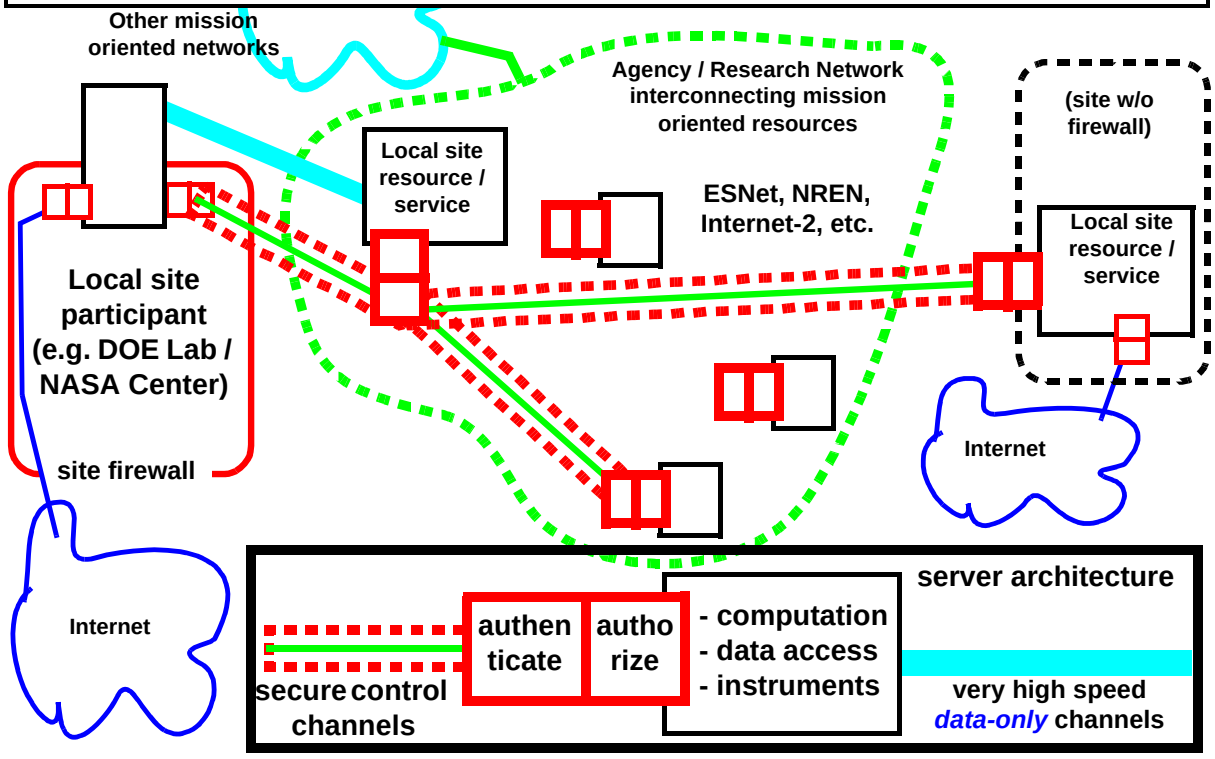
  - ***computing and on-line storage*** **resources (initially drawn from the R&D community)**
  - ***archival storage*** **resources that are *uniformly* accessible from all DSG testbed systems**
  - ***collaboration and instrument systems***
  - ***end-to-end network interconnects*** **of at least 100 mbit/s between resources + QoS**

Establish a security model

A security model for the DSG testbed must address:

- **cyber risk mitigation and cross-site integrity**
- **control channel integrity and confidentiality**
- **optional data channel integrity and confidentiality**
- **identity management, authentication, and single identity sign-on w/o clear text passwords**
- **authorization via policy-based access control**
- **infrastructure assurance**

Security Model: All Grid services command and control functions are transported over an encrypted channel, after the client/user is authenticated and authorized. This compartmentalizes all servers: If multiple servers are involved in a distributed system, then each reauthorizes connections through the use of cryptographic proxies or active re-authentication.



Expected Outcomes

Near-term

- ***DOE Science Grid Testbed: An operational, prototype Grid*** **environment incorporating computing, data and instrument resources at multiple DOE sites**
- ***Collaboration initially between LBNL, ANL, and SNL*****, with others expected to join (e.g. LANL/ACL, PNNL, ORNL, etc.)**

- ***Support for advanced data and computation applications*** **and environments (e.g. NGI++, GC++, Physics Grid, ...) – several "benchmark" applications operating across DSG testbed**
- ***Collaboration services and tools*****, and remote instrument systems integrated with DSG**
- ***Identification of weaknesses and missing capabilities in the current Grid components*** **with respect to DOE applications**

Expected Outcomes

- ***Support for Grid R&D and integration of current R&D***
  - **Grid Information Service functionality, performance, scalability, and robustness (ESNet, DSD, and IPG)**
  - **policy based authorization and access control (DSD)**
  - **network QoS and bandwidth reservation (ESNet, DSD, ANL)**
  - **high bandwidth, wide area network transport (DSD)**

Expected Outcomes - Grid R&D

- **Grid enabled collaboration and visualization (DSD, ESNet, and NERSC)**
- **distributed network cache as a Grid service (DSD)**
- **network aware middleware (DSD)**
- **resource brokering (ANL)**
- **configurable and secure monitoring and management of distributed components (DSD)**
- **secure group communication ("reliable multicast") (DSD and UCI)**

- ***Support of application R&D for using the dynamically assembled, widely distributed systems provided by Grids***
- ***Experience on how - within DOE - to provide persistent and supported Grid infrastructure*** **capable of routinely and uniformly accessing and sharing distributed resources**

Long-term

- ***A Production DOE Science Grid*** **that provides uniform and routine access to widely distributed DOE computing, data handling, instrumentation, and human resources**
- ***Enabling the DOE scientific community to routinely address larger scale, more diverse, and more transient problems than is possible today***

- **Specialized Grid R&D testbeds:**
  - ***Sys*** **testbed: Grid system software development**
  - ***HDR*** **testbed: high data-rate distributed resources**
    - **high-speed end-to-end**
    - **high data-rate services including data archives and instrument systems**
    - **test applications**
  - ***Sec*** **testbed: security and infrastructure protection**

Roadmap for the DSG Testbed

time	activity	outcome
FY00/01	ESNet and DSD prototyping basic Grid services in DSG testbed	deployed and supported DSG testbed
	participation/testing by NERSC center staff	
	incorporate computing and data resources into DSG testbed	
	incorporate current R&D	
FY01/02	ESNet supporting basic Grid services in DSG testbed	
	applications using DSG	
	establish NERSC internal Grid	prototype-production DSG
	incorporate some NERSC resources into prototype-production DSG	
	establish Grid R&D agenda	
FY02/03	operational, baseline DSG	production DSG providing access to significant DOE resources
	access to NERSC Grid from DSG	

Tasks for a minimal persistent infrastructure: DSG testbed startup tasks (FY00, Q3-4 / FY01, Q1-2-3)

task	who	effort (FTE) existing[a]	new	total
1. Establish the DSG testbed Grid Information Service	ESNet + DSD[b]	1.0	0.0	1.0
2. Establish DSG testbed X.509 Certification Authority and certificate server	ESNet + DSD	0.5	0.5	0.5
3. Identify computing resources for the initial DSG testbed multi-Lab testbed	DSD	0.5	0.5	1.0
	ANL	ANL		
	other sites			
4. Deploy Globus across DSG testbed	DSD	1.0	1.0	2.0
	ANL	ANL		
	other sites			
5. Establish networking for the DSG testbed	ESNet + site LAN groups	largely part of NGI testbed work		

task	who	effort (FTE) existing[a]	new	total
6. Provide global queuing and user-level queue management capability on top of Globus (w/NASA)	HPCRD[c]	0.0	1.0	1.0
7. Identify DSG testbed strategy for uniform access for archival and published data: • GASS and DPSS integrated with HPSS • SDSC's Metadata Catalogue (MCAT) and the Storage Resource Broker (SRB)	DSD + NERSC	1.0	1.0	2.0
	ANL	ANL		
8. Facilitate use of DSG testbed by applications	DSD	0.5		0.5
	HPCRD/Vis Group	0.0	1.0	1.0
	NERSC	1.0	0.0	1.0
	ANL	ANL		
	existing NGI apps			
	other sites			

task	who	effort (FTE)		
		existing[a]	new	total
9. Incorporate collaboration tools into DSG	DSD	0.5	0.5	1.0
	ANL	ANL		
10. DSG network testbed + QoS	ESNet	1.0	0.0	1.0
	DSD	1.0	0.0	1.0
	ANL	ANL		
11. Define and implement the security model	DSD	0.5	0.5	1.0
12. Incorporate an instrument system into DSG testbed	NGI project? ALS BL-7? APS?	?		
subtotals		8.5	6.0	14.5

a. In some cases this assumes that current NGI funding, or equivalent, continues.

b. DSD = LBNL Distributed Systems Dept.

c. HPCRD = LBNL High Performance Computing Research Dept.

DSG operational infrastructure tasks(FY02, Q4 / FY03, Q1-2-3)

task	who	effort (FTE)		
		existing	new	total
13. Provide for maintenance for Grid Information Service / MDS database & CA	ESNet		1.0	1.0
14. Provide for DSG/Globus system administration	DSD / NERSC	0.5 (from T4[a])	NERSC[b]	1.5
15. Programming and program execution support	DSD / NERSC	1.0 (from T8[a])	NERSC	1.0
16. Develop automatic monitoring of DSG components	DSD	0.0	0.5	1.0
17. Job tracking and monitoring	NERSC		NERSC	0.5
18. Security programming standards and enforcement mechanisms	DSD	0.5	NERSC	1.0
19. Trouble ticket system	NERSC			

task	who	effort (FTE)		
20. Provide documentation of DSG specific topics	all, working with NERSC			
21. Provide for user services	NERSC	0.0	NERSC	1.0
subtotals		2.0	1.5	7.0

a. Transition from development to prototype-production

b. It is expected that some aspects of the DOE Science Grid Development effort (under a separate FWP) will be deployable to a large scale facility in 2002. When this happens, explicit resources will have to be allocated to the testing, integration, support and services of the tools made available to NERSC clients. It will also require working with vendors to incorporate key software functionality on advanced computational and storage systems. In order to do this, an increase of 2 FTE will be required to augment the NERSC Facility staff.

Prototype production DSG tasks (FY03/04)

task	who	effort
22. Operational support	NERSC & ESNet	TBD
23. Provide for account management (automated generation and maintenance mechanisms)	NERSC	TBD
24. Develop allocation management and accounting tools	NERSC	TBD
25.System testing: Verification suites, benchmarks, and reliability/sensitivity analysis for DSG (both static and dynamic)	DSD & NERSC	TBD

References and Acronyms

[1] Globus is a middleware system that provides a suite of services designed to support high performance, distributed applications. Globus provides:

- Resource Management: Components that provide standardized interfaces to various local resource management systems (GRAM) manage allocation of collections of resources (DUROC). All Globus resource management tools are tied together by a uniform resource specification language (RSL).
- Remote Access: Components that enable remote access to files (GASS and RIO) and executables (GEM).
- Security: Support for single sign-on, authentication, and authorization within the Globus system (GSI) and (experimentally) authorization (GAA).
- Fault Detection: Basic support for building fault detection and recovery into Globus applications.
- Information Infrastructure: Global access to information about the state and configuration of system components of an application (MDS).
- Grid programming services: Support writing parallel-distributed programs (MPICH-G), monitoring (HBM), etc.

www.globus.org provides full information about the Globus system.

[2] *The Grid: Blueprint for a New Computing Infrastructure*, edited by Ian Foster and Carl Kesselman. Morgan Kaufmann, Pub. August 1998. ISBN 1-55860-475-8. http://www.mkp.com/books_catalog/1-55860-475-8.asp

[3] "Grids as Production Computing Environments: The Engineering Aspects of NASA's Information Power Grid," William E. Johnston, Dennis Gannon, and Bill Nitzberg. Eighth IEEE International Symposium on High Performance Distributed Computing, Aug. 3-6, 1999, Redondo Beach, California. (Available at http://www.nas.nasa.gov/~wej/IPG)

[4] See www.nas.nasa.gov/IPG for project information and pointers.

[5] See http://www-itg.lbl.gov/NGI/ for project information and pointers.

[6] Tierney, B. Lee, J., Crowley, B., Holding, M., Hylton, J., Drake, F., "A Network-Aware Distributed Storage Cache for Data Intensive Environments", Proceeding of IEEE High Performance Distributed Computing conference (HPDC-8), August 1999.

[7] "Real-Time Generation and Cataloguing of Large Data-Objects in Widely Distributed Environments," W. Johnston, Jin G., C. Larsen, J. Lee, G. Hoo, M. Thompson, and B. Tierney (LBNL) and J. Terdiman (Kaiser Permanente Division of Research). Invited paper, International Journal of Digital Libraries - Special Issue on "Digital Libraries in Medicine". May, 1998. http://www-itg.lbl.gov/WALDO/

[8] MAGIC: "The MAGIC Gigabit Network." See: http://www.magic.net

[9] TerraVision-2: VRML based data fusion and browsing - www.ai.sri.com/TerraVision

[10] A collaborative effort to enable desktop access to remote resources including, supercomputers, network of workstations, smart instruments, data resources, and more - computingportals.org

[11] The Clipper Project: Computational Grids providing middleware that supports applications requiring configurable, distributed, high-performance computing and data resources. See http://www-itg.lbl.gov/~johnston/Clipper

[12] The Grid Forum (www.gridforum.org) is an informal consortium of institutions and individuals working on wide area computing and computational Grids.

[13] The Portable Batch System (PBS) is a batch queueing system developed at NAS. PBS implements the POSIX standard, and operates on networked, multi-platform UNIX environments, including heterogeneous clusters of workstations, supercomputers, and massively parallel systems. PBS is the basis of the IPG global queuing system work. http://parallel.nas.nasa.gov/Parallel/PBS/

[14] Storage Resource Broker (SRB) provides uniform access mechanism to diverse and distributed data sources. http://www.sdsc.edu/MDAS/

[15] Condor is a High Throughput Computing environment that can manage very large collections of distributively owned workstations. http://www.cs.wisc.edu/condor/

[16] SCIRun is a scientific programming environment that allows the interactive construction, debugging and steering of large-scale scientific computations. http://www.cs.utah.edu/~sci/software/

[17] WebFlow - A prototype visual graph based dataflow environment, WebFlow, uses the mesh of Java Web Servers as a control and coordination middleware, WebVM. See http://iwt.npac.syr.edu/projects/webflow/index.htm